Psychological Medicine (2013), **43**, 2649–2656. © Cambridge University Press 2013
doi:10.1017/S0033291713000342

ORIGINAL ARTICLE

Use of self-administered instruments to assess psychiatric disorders in older people: validity of the General Health Questionnaire, the Center for Epidemiologic Studies Depression Scale and the self-completion version of the revised Clinical Interview Schedule

J. Head[1]*, S. A. Stansfeld[2], K. P. Ebmeier[3], J. R. Geddes[3], C. L. Allan[3], G. Lewis[4] and M. Kivimäki[1]

[1] *Department of Epidemiology and Public Health, University College London Medical School, London, UK*
[2] *Centre for Psychiatry, Wolfson Institute of Preventive Medicine, Queen Mary's School of Medicine and Dentistry, London, UK*
[3] *Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK*
[4] *Academic Unit of Psychiatry, University of Bristol, Bristol, UK*

Background. Diagnosis of depressive disorder using interviewer-administered instruments is expensive and frequently impractical in large epidemiological surveys. The aim of this study was to assess the validity of three self-completion measures of depressive disorder and other psychiatric disorders in older people against an interviewer-administered instrument.

Method. A random sample stratified by sex, age and social position was selected from the Whitehall II study participants. This sample was supplemented by inclusion of depressed Whitehall II participants. Depressive disorder and other mental disorders were assessed by the interviewer-administered structured revised Clinical Interview Schedule (CIS-R) in 277 participants aged 58–80 years. Participants also completed a computerized self-completion version of the CIS-R in addition to the General Health Questionnaire (GHQ) and the Center for Epidemiologic Studies Depression Scale (CES-D).

Results. The mean total score was similar for the interviewer-administered (4.43) and self-completion (4.35) versions of the CIS-R [95% confidence interval (CI) for difference −0.31 to 0.16]. Differences were not related to sex, age, social position or presence of chronic physical illness. Sensitivity/specificity of self-completion CIS-R was 74%/98% for any mental disorder and 75%/98% for depressive episode. The corresponding figures were 86%/87% and 78%/83% for GHQ and 77%/89% and 89%/86% for CES-D.

Conclusions. The self-completion computerized version of the CIS-R is feasible and has good validity as a measure of any mental disorder and depression in people aged ⩾60 years. GHQ and CES-D also have good criterion validity as measures of any mental disorder and depressive disorder respectively.

Received 18 April 2012; Revised 24 January 2013; Accepted 31 January 2013; First published online 14 March 2013

Key words: Anxiety, common mental disorder, depressive disorder, mental health, method comparison.

Introduction

Structured diagnostic interviews, such as the Composite International Diagnostic Interview (CIDI; Wittchen, 1994) and the revised Clinical Interview Schedule (CIS-R; Lewis *et al.* 1992), are considered by many researchers to be the most valid and reliable methods for the assessment of mental disorders in populations according to diagnostic criteria (ICD-10 or DSM-IV). The CIS-R has been widely used in the UK (Brugha *et al.* 1999) whereas the CIDI has been more commonly used in the USA (Haro *et al.* 2006). In comparisons with semi-structured clinical evaluations, the CIS-R has been shown to be a valid measure of mental disorders (Patton *et al.* 1999; Jordanova *et al.* 2004; Brugha *et al.* 2005; Pez *et al.* 2010).

However, structured interviews such as the CIDI and the CIS-R may be expensive and impractical to use in large, epidemiological studies. Large-scale surveys have therefore often relied on self-administered instruments to identify psychiatric illness and morbidity,

* Address for correspondence: Ms. J. Head, Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK.
(Email: j.head@ucl.ac.uk)

despite concerns about the validity and reliability of these measures. Although some studies have demonstrated that self-administered instruments are valid in younger and middle-aged adults (Goldberg & Williams, 1988; Stansfeld & Marmot, 1992) and have compared self-completion and interviewer versions of either the CIS-R or the CIDI (Lewis *et al.* 1988; Lewis, 1994; Peters *et al.* 1998), few studies have investigated their validity in older populations.

In this study, we tested whether a computerized self-completion version of the CIS-R (Lewis *et al.* 1988) was a feasible and valid instrument for identifying mental disorders in older adults by comparing results with the interview-administered CIS-R, considered to be the reference standard in this study. In addition, we examined the sensitivity and specificity of two commonly used self-completion questionnaires, the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) and the General Health Questionnaire (GHQ; Goldberg, 1972), as measures of psychiatric disorders in a UK population aged 58–80 years.

Method

The Whitehall II study

The Whitehall II study is a cohort of 10 308 originally London-based civil servants (6895 men and 3413 women) aged between 35 and 55 years in 20 London-based civil service departments, established between 1985 and 1988 (phase 1) (Marmot *et al.* 1991). The validation study reported in this paper was conducted at phase 10 of the Whitehall II study in 2011. The main aims of phase 10 were to (1) validate self-completion measures of psychiatric morbidity, in addition to several other screening measures, in older people and (2) to invite a subsample of participants to take part in a neuroimaging study of late-onset depression cases and never-depressed controls.

Study sample

The sample (phase 10) was selected from the Whitehall II cohort. We drew a random sample of 255 persons, stratified by sex, age and social position (most recent employment grade) from among the 5390 cohort members who attended the phase 9 follow-up examination in 2008–2009. To obtain a sufficient number of depressed adults, we supplemented this sample by inclusion of all participants with evidence of late-onset depressive symptoms in the 2008–2009 follow-up.

Of the 5390 who attended the phase 9 screening examination, 88 participants were classified as having late-onset depressive symptoms; as six of these 88 were already selected in the random sample of 255, this gave a total supplemented sample of 337. Three of the 337 were living overseas and therefore were not invited to participate; a further four people died before being contacted. Thus, 330 people were eligible and were invited to participate at phase 10; of these, 277 took part (response rate 84%).

Study procedures and measures

Self-completion questionnaires, including the CES-D and the GHQ, were sent out in December 2010 along with invitation letters to attend a screening clinic. Participants were asked to bring along their completed questionnaires to hand in at the clinic. According to the recorded date of questionnaire completion, the majority of participants completed their postal questionnaires shortly before their screening clinic appointment (median 2 days apart, 87% less than 30 days apart). Between 31 January 2011 and 14 March 2011, participants attended screening where they completed both the interviewer-administered and the computerized self-completion versions of the CIS-R. We allocated participants randomly to complete either the interviewer version first or the computerized self-completion version first. A potential limitation of our study is that, to reduce respondent burden, both versions of the CIS-R were administered on the same day. However, participants were administered the other phase-10 measures in between the first and second CIS-R versions to reduce the risk that the respondent recalled their answer to the first version. Participants were offered tea and biscuits at the end of the phase-10 screening but no financial incentives were offered for participation. Ethical approval for the Whitehall II study was obtained from the University College London Medical School committee on the ethics of human research, and all participants gave informed written consent.

The CIS-R is a structured diagnostic interview for common mental disorders, formerly neurotic disorders (Lewis *et al.* 1992), but because of the structured nature of the questions and responses in this measure, a computerized self-completion version is also available (Lewis *et al.* 1988). Both versions generate scores on 14 psychiatric symptoms (listed in Table 1), a total score and diagnoses of depressive and other common mental disorders based on the ICD-10 (diagnoses listed in Table 2), thus providing measures of severity and also presence or absence of mental disorders.

A CIS-R total score $\geqslant$12 was used to define cases with any common mental disorder (Lewis *et al.* 1992). The wording of the questions and responses was the same in the computerized self-completion and interviewer-administered versions but interviewers used show cards listing response options for questions that were sensitive or had several possible responses.

https://doi.org/10.1017/S0033291713000342

Table 1. *Prevalence of mental disorders*

	Prevalence (n=274)		Weighted prevalence (n=274)	Prevalence in cohort subsample (n=214)
	No. of cases	% (95% CI)	% (95% CI)	% (95% CI)
CIS-R Any mental disorder				
Interview	27	9.9 (6.3–13.4)	5.9 (3.5–9.8)	5.6 (2.5–8.7)
Self-completion	24	8.8 (5.4–12.1)	4.9 (2.8–8.6)	4.7 (1.8–7.5)
CIS-R Specific disorders[a] (ICD-10 code)				
Depressive episode (F32)				
Interview	12	4.4 (1.9–6.8)	3.8 (1.9–7.3)	3.7 (1.2–6.3)
Self-completion	15	5.5 (2.8–8.2)	3.4 (1.7–6.8)	3.3 (0.9–5.7)
Mixed anxiety and depressive disorder (F41.2)				
Interview	9	3.3 (1.2–5.4)	1.5 (0.5–4.2)	1.4 (0.1–3.0)
Self-completion	6	2.2 (0.4–3.9)	1.0 (0.3–3.6)	0.9 (0.1–2.2)
Generalized anxiety disorder (F41.1)				
Interview	16	5.8 (3.0–8.6)	4.7 (2.6–8.5)	4.7 (1.8–7.5)
Self-completion	11	4.0 (1.7–6.3)	2.4 (1.1–5.5)	2.3 (0.3–4.4)
All phobias (F40)				
Interview	5	1.8 (0.2–3.4)	1.0 (0.3–3.6)	0.9 (0.1–2.2)
Self-completion	9	3.3 (1.2–5.4)	1.5 (0.5–4.2)	1.4 (0.1–3.0)
		(n=256)	(n=256)	(n=204)
CES-D case	43	16.8 (12.2–21.4)	10.7 (7.2–15.6)	10.3 (6.1–14.5)
GHQ case	49	19.1 (14.3–24.0)	15.0 (10.7–20.5)	14.7 (9.8–19.6)
GHQ depression case	21	8.2 (4.8–11.6)	6.0 (3.5–10.2)	5.9 (2.6–9.1)

CIS-R, Clinical Interview Schedule Revised; CI, confidence interval; CES-D, Center for Epidemiologic Studies Depression Scale; GHQ, General Health Questionnaire.

[a] There were two cases of panic disorder (ICD-10 F41.0) and no cases of obsessive–compulsive disorder (ICD-10 F42).

Both versions were administered using the program PROQSY (Lewis *et al.* 1988). A laptop computer was used for the interviewer version with interviewers reading questions from the screen and entering responses directly. A desktop computer was used for the self-completion version. Start and end times were recorded so that completion times could be compared. Interviewers were given 2 hours of training in the use of the CIS-R, which included a practice session. All interviewers were given a written protocol to follow and had the opportunity for further practice interviews during the pilot of the phase-10 data collection. Less than 1% of participants were given help with the computerized version because of eyesight problems or other problems with using a computer.

The 20-item CES-D is a short self-report questionnaire designed to measure depressive symptoms in the general population (Radloff, 1977). Participants were asked to score the frequency of occurrence of specific symptoms during the previous week on a four-point scale, where 0='less than 1 day', 1='1–2 days', 2='3–4 days' and 3='5–7 days'. These were summed to yield a total score between 0 and 60. Participants scoring $\geqslant$16 were categorized as cases of CES-D depression (Stansfeld *et al.* 2008). The CES-D was included at phases 7, 9 and 10.

The 30-item GHQ is a well-established screening questionnaire for common mental disorder, suitable for use in general population samples (Goldberg, 1972). The GHQ was included in all study phases 1–10 with the exception of phase 4. At phase 1 of the study, this was validated against the CIS in a subsample and, on the basis of receiver operating characteristic (ROC) analysis, those scoring $\geqslant$5 were deemed GHQ cases (Stansfeld & Marmot, 1992). A four-item depression subscale (Cronbach's α=0.88) was identified from the 30-item GHQ on the basis of factor analysis and comparison with the items of the depression subscale of the 28-item GHQ (Goldberg & Hillier, 1979). A total depression score (ranging from 0 to 12) was derived by summing responses to these four items using Likert scoring (0 to 3) for each item. Participants scoring $\geqslant$2 were categorized as cases of GHQ depression (Stansfeld *et al.* 1998).

A measure of early-onset depressive symptoms was derived from GHQ measures at phases 1–9 and

https://doi.org/10.1017/S0033291713000342

Table 2. *Agreement between self-completion and interviewer versions for the total CIS-R score and symptom scores (n=274)*

	Mean self-completion	Mean interview	Difference in means	95% CI for difference in means	Weighted κ
Total CIS-R score[a]	4.35	4.43	−0.08	−0.31 to 0.16	0.94
Symptom scores[b]					
Somatic	0.23	0.14	0.08	−0.01 to 0.18	0.31
Fatigue	0.51	0.59	−0.08	−0.16 to −0.01	0.81
Concentration and forgetfulness	0.31	0.31	0	−0.05 to 0.05	0.77
Sleep	0.86	0.91	−0.05	−0.12 to 0.02	0.87
Irritability	0.36	0.32	0.04	−0.01 to 0.09	0.78
Worry over physical health	0.26	0.30	−0.04	−0.09 to 0.01	0.76
Depression	0.23	0.19	0.04	−0.01 to 0.10	0.71
Depressive ideas	0.20	0.15	0.05	0.01 to 0.10	0.84
Worry	0.50	0.61	−0.11	−0.18 to 0.03	0.79
Anxiety	0.27	0.30	−0.04	−0.11 to 0.04	0.62
Phobias	0.14	0.10	0.03	−0.01 to 0.07	0.80
Panic	0.03	0.03	0	−0.01 to 0.01	0.94
Compulsions	0.17	0.24	−0.07	−0.13 to −0.01	0.69
Obsessions	0.29	0.24	0.05	−0.03 to 0.13	0.56

CIS-R, Clinical Interview Schedule Revised; CI, confidence interval.

[a] Total score ranges from 0 to 57.

[b] Symptom scores range from 0 to 4 (depressive ideas symptom score 0 to 5).

defined as two or more reports of GHQ caseness and/or two or more reports of GHQ depression subscale caseness before age 60. Late-onset depression was defined as being a CES-D case at phase 9 AND having no early-onset depressive symptoms AND no prevalent serious chronic conditions (coronary heart disease, cancer, stroke).

Statistical analysis

For each mental disorder we computed estimates of raw prevalence, weighted prevalence to adjust for oversampling of depressed cases and prevalence in the randomly selected cohort subsample. Differences in prevalence estimates between the self-completion measures and interviewer CIS-R were tested using McNemar's χ^2 test. Differences in mean total scores and specific symptom scores between the self-completed and interviewer CIS-R were examined with the paired *t* test. The agreement of scores between the two versions was assessed with the weighted κ statistic. Linear regression with difference in CIS-R score between the two versions was used to test for evidence that differences in method of administration were related to age, sex, employment grade or presence of chronic physical illness. We performed ROC analysis to compute estimates of sensitivity, specificity, positive predictive value (+PV), negative predictive value (–PV) and area under the ROC curve (AUC) for all self-completion measures of any mental disorder and specific mental disorders using the interviewer-administered CIS-R as the criterion. Based on published guidelines, we considered AUC values $\geqslant$0.90 to indicate excellent validity and values $\geqslant$0.80 but <0.90 to indicate good validity (Metz, 1978). We checked the cut-off points of scores $\geqslant$16 and scores $\geqslant$5 used to define CES-D and GHQ cases respectively by ROC analyses. Analyses were performed using Stata version 12 (StataCorp, USA).

Results

Of the 330 persons invited, 277 attended the examination (response rate 84%) and 274 had complete data on both interviewer-administered and self-completion versions of the CIS-R. The mean age was 69.1 (S.D. = 5.8) years for participants allocated to the self-completion CIS-R version first and 68.3 (S.D. = 6.2) for participants allocated to the self-completion CIS-R version second. Among participants allocated to the self-completion version first (second), 31% (28%) were female; the most recent employment grade was high for 42% (47%), middle for 45% (41%) and low for 13% (13%); the proportion classified as GHQ cases was 21% (21%) and the proportion classified as CES-D cases was 17% (16%). Similarly, CIS-R mean total scores did not differ significantly according to order of administration of the two CIS-R versions.

https://doi.org/10.1017/S0033291713000342

Table 3. *Sensitivity and specificity for the self-completion CIS-R assessment of mental disorders with the interviewer-administered CIS-R as the criterion (n=274)*

	No. of cases Interviewer/ self-completion	*p* value[a]	Sensitivity % (95% CI)	Specificity % (95% CI)	+PV	−PV	+LR	−LR	AUC (95% CI)
Any mental disorder	27/24	0.37	74.1 (0.53–0.88)	98.4 (0.96–0.99)	0.83	−0.97	45.74	0.26	0.86 (0.78–0.95)
Specific disorders (ICD-10 code)									
Depressive episode (F32)	12/15	0.32	75.0 (0.43–0.97)	97.7 (0.95–0.99)	0.60	0.99	32.75	0.26	0.86 (0.74–0.99)
Mixed anxiety and depressive disorder (F41.2)	9/6	0.41	11.1 (0.01–0.49)	98.1 (0.95–0.99)	0.17	0.97	5.89	0.91	0.55 (0.44–0.66)
Generalized anxiety disorder (F41.1)	16/11	0.20	37.5 (0.16–0.64)	98.1 (0.95–0.99)	0.56	0.96	19.35	0.64	0.68 (0.56–0.80)
All phobias (F40)	5/9	0.10	80.0 (0.30–0.99)	98.1 (0.95–0.99)	0.44	0.99	43.04	0.20	0.89 (0.69–0.99)

CIS-R, Clinical Interview Schedule Revised; CI, confidence interval; +PV, positive predictive value; −PV, negative predictive value; +LR, positive likelihood ratio; −LR, negative likelihood ratio; AUC, area under the receiver operating characteristic (ROC) curve.

[a] McNemar's χ^2 test for difference in prevalence between interviewer and self-completion versions.

Table 1 presents the prevalence for each of the mental health measures. Based on the interviewer-administered CIS-R, 27 participants were diagnosed as having any mental disorder. The numbers of participants diagnosed as having specific disorders were: 12 depressive episode; nine mixed anxiety and depressive disorder; 16 generalized anxiety disorder; five phobia; and two panic disorder. No participants were diagnosed with obsessive–compulsive disorder.

Validity of self-completion CIS-R

Table 2 shows the mean value for the total CIS-R score and each of the 14 symptom scores. The mean difference in the total score between self-completion CIS-R and interviewer CIS-R was small, the mean scores on the two versions were 4.35 and 4.43 respectively [95% confidence interval (CI) for difference in means −0.31 to 0.16, $p=0.26$, paired t test]. For 12 of the 14 symptom scores, differences in symptom scores did not differ according to method of administration. Differences for both fatigue and compulsions were statistically significant, with slightly lower scores on the self-completion version than the interviewer version. In a linear regression model, the difference in total CIS-R score between the two versions was not related to age, sex, social position or presence of chronic physical illness.

Table 3 presents sensitivity and specificity figures for the self-completion CIS-R measures of any mental disorder and specific mental disorders. The sensitivity for any mental disorder was 74.1% and specificity 98.4%. The corresponding figures for depressive episode were 75.0% and 97.7% respectively. The self-completion CIS-R was also a sensitive and specific measure of all phobias (80%/98.1%), but its sensitivity was low for mixed anxiety and depressive disorder and for generalized anxiety disorder. The specificity (>97%) was very high for all diagnostic categories.

Validity of the CES-D and GHQ

Table 4 shows that the CES-D is a sensitive and specific measure of any mental disorder (sensitivity/specificity 77%/89%) and depressive episode (sensitivity/specificity 89%/86%). This is also the case for the GHQ caseness (86%/87% for any mental disorder; 78%/83% for depressive episode). By contrast, the GHQ depression measure constructed from four items of the 30-item GHQ was not a sensitive measure for depressive episode, although the ROC analysis indicated that sensitivity for depressive episode was somewhat improved for a cut-point ⩾2 (sensitivity/specificity

https://doi.org/10.1017/S0033291713000342

Table 4. *Sensitivity and specificity for the self-completion CES-D and the GHQ as measures of any mental disorder and depressive episode with the interviewer-administered CIS-R as the criterion (n=256)*

	CIS-R interviewer version						
	Sensitivity (%)	Specificity (%)	+PV	−PV	+LR	−LR	AUC (95% CI)
Any mental disorder							
CES-D case	77.3 (0.54–0.91)	88.9 (0.84–0.92)	0.39	0.98	6.95	0.26	0.83 (0.74–0.92)
GHQ case	86.4 (0.64–0.96)	87.2 (0.82–0.91)	0.39	0.99	6.74	0.16	0.87 (0.79–0.94)
GHQ depression case	36.4 (0.18–0.59)	94.4 (0.80–0.97)	0.38	0.94	6.55	0.67	0.65 (0.55–0.76)
Depressive episode							
CES-D case	88.9 (0.51–0.99)	85.8 (0.81–0.90)	0.19	0.99	6.27	0.13	0.87 (0.76–0.98)
GHQ case	77.8 (0.40–0.96)	83.0 (0.78–0.87)	0.14	0.99	4.57	0.27	0.80 (0.66–0.95)
GHQ depression case	44.4 (0.15–0.77)	93.1 (0.89–0.96)	0.19	0.98	6.46	0.60	0.69 (0.51–0.86)

CES-D, Center for Epidemiologic Studies Depression Scale; GHQ, General Health Questionnaire; CIS-R, Clinical Interview Schedule Revised; +PV, positive predictive value; −PV, negative predictive value; +LR, positive likelihood ratio; −LR, negative likelihood ratio; AUC, area under the receiver operating characteristic (ROC) curve; CI, confidence interval.

56%/90%) in place of the cut-point ⩾3 used in earlier studies (sensitivity/specificity 44%/93%).

Discussion

Data from men and women aged 58–80 years show reasonably high sensitivity and specificity, varying between 74% and 98%, for the CES-D, the 30-item GHQ and the computerized self-completion version of the CIS-R as measures of any mental disorder and depressive episode. The computerized self-completion CIS-R was additionally a sensitive and specific measure of phobias and accurately detected symptom severity in 12 specific psychiatric symptoms. These findings suggest that several self-administered instruments, with reasonable criterion validity, may be used to screen for common mental disorders and depression in populations aged ⩾60 years. Furthermore, the mean total score from the computerized self-completion version and the structured interview version were very similar.

An earlier comparison of the computerized self-completion version of the CIS-R against the structured psychiatric interview in this population when they were aged 35–55 years showed slightly higher sensitivity (82%) and lower specificity (84%) (Lewis *et al.* 1988). Previous studies on this measure have shown good agreement in severity score and case definition for any psychiatric disorder in primary care and occupational settings but these studies did not examine agreement for specific ICD-10 disorders such as depressive episode (Lewis *et al.* 1988; Lewis, 1994). We found that symptom scores were significantly lower on the self-completion version than the interviewer version for both fatigue and compulsions. This is in contrast to an earlier study where the only significant difference in the 14 symptom scores was for sleep symptoms (Lewis, 1994). It is possible that these findings are due to chance.

According to a review of 28 studies, previous investigations on the CES-D and GHQ have reported validity estimates comparable to those we observed (Williams *et al.* 2002). Our current findings are also in agreement with those obtained over 20 years ago for this cohort. At the baseline of the Whitehall II study when the participants were aged 35–55 years, the sensitivity of the GHQ against the CIS was 73% although specificity was slightly worse at 78% (Stansfeld & Marmot, 1992). In a vulnerable, very old population living in residential homes in The Netherlands, sensitivity for CES-D for depressive and/or anxiety disorders exceeded 80% but specificity was lower, at 61% (Dozeman *et al.* 2011). Among postpartum women, a 60% sensitivity and 90% specificity was observed for the CES-D (Boyd *et al.* 2005). However, the validity of the CES-D has been lower in some (Klinkman *et al.* 1997; Thomas *et al.* 2001) but not all clinical samples (Stahl *et al.* 2008).

Limitations and strengths of the study

A limitation of this study is that participants were recruited from an occupational cohort so our findings may not apply to people who have not had paid employment. Our sample was relatively healthy and consisted of people able to travel to our London clinic. Estimates of sensitivity were imprecise for specific anxiety disorders because of the small number of people diagnosed with these disorders in this sample. We considered the interviewer-administered version

https://doi.org/10.1017/S0033291713000342

to be the 'gold standard' criterion although this is somewhat arbitrary as it is possible that people may be more likely to under-report symptoms in an interviewer-administered version than in a self-completion version. Given this limitation, our study could alternatively be described as a reliability, method-comparison or concordance study. Furthermore, the GHQ and the CES-D self-completion questionnaire were posted to participants so that differences between the GHQ/CES-D and the CIS-R may be attributable not only to the instrument but also to the mode of administration and setting, such as completion at home rather than in a clinic. A further limitation is that, although the majority of participants completed their postal questionnaires shortly before their screening clinic appointment (median 2 days apart, 87% within 1 month), the gap of more than a month for some participants may mean that the results were influenced by changes in symptoms. It is possible that this partially accounted for our results showing that sensitivity was poor for both mixed anxiety and depressive disorder and for generalized anxiety disorder.

The strengths of this study are that our sample was selected randomly from a large cohort study, and was large enough to demonstrate that the similar severity scores obtained from the two methods of administration were consistent for men and women, across age groups, for different employment grades and for people with and without a chronic physical illness. Additionally, we demonstrated that it is feasible to use a computerized self-completion version in studies of older participants as response rates were identical for the two versions.

An advantage of self-completion instruments is that they are less expensive to administer than interviewer instruments. At the time of writing this paper, more than 1500 participants had been screened in the sixth medical examination of the Whitehall II study. Respondents attended the clinic where physiological measures, blood tests, cognitive function and the self-completion version of the CIS-R were administered. A member of the clinic staff introduced the respondent to the self-completion computerized CIS-R version. This took no more than a few minutes. Several computers were available in a quiet room so that up to six respondents could complete the CIS-R at any one time. We estimate that using the self-administered CIS-R procedure reduced staff costs by at least 60% compared to using the interviewer version, where it would be necessary to schedule appointments about 30 to 45 minutes apart. Based on preliminary data from the first 1500 participants at phase 11, 0.5% were given reading glasses and 0.5% were helped by clinic staff because of poor eyesight or physical difficulty using a computer.

Implications

Taken together, these findings suggest that the computerized self-completion CIS-R provides a feasible and less expensive alternative to the interviewer-administered CIS-R to identify any common mental disorder and depressive episode according to ICD-10. The CES-D and 30-item GHQ also have reasonable criterion validity as measures of common mental disorders and depression.

Acknowledgements

We thank all participating civil service departments and their welfare, personnel and establishment officers; the Occupational Health and Safety Agency; the Council of Civil Service Unions; all participating civil servants in the Whitehall II study; and all members of the Whitehall II study team.

The Whitehall II study has been supported by grants from the Medical Research Council (MRC) G8802774; the British Heart Foundation; the Health and Safety Executive; the Department of Health; the National Heart, Lung, and Blood Institute (R01HL036310); the National Institute on Aging, National Institutes of Health (NIH) (R01AG013196 and R01AG034454); and the Agency for Health Care Policy and Research (grant HS06516). J. Head was supported in part by the National Institute on Aging, NIH (R01AG013196). M. Kivimäki was supported by the MRC, the Academy of Finland, and a professorial fellowship from the Economic and Social Research Council, UK. C. Allan received support from Oxford University Clinical Academic Graduate School (OUCAGS).

Declaration of Interest

None.

References

Boyd RC, Le HN, Somberg R (2005). Review of screening instruments for postpartum depression. *Archives of Women's Mental Health* **8**, 141–153.

Brugha TS, Bebbington PE, Jenkins R (1999). A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychological Medicine* **29**, 1013–1020.

Brugha TS, Meltzer H, Jenkins R, Bebbington PE, Taub NA (2005). Comparison of the CIS-R and CIDI lay diagnostic interviews for anxiety and depressive disorders. *Psychological Medicine* **35**, 1089–1091.

Dozeman E, van Schaik DJ, van Marwijk HW, Stek ML, van der Horst HE, Beekman AT (2011). The Center for Epidemiological Studies Depression Scale (CES-D) is an

https://doi.org/10.1017/S0033291713000342

adequate screening instrument for depressive and anxiety disorders in a very old population living in residential homes. *International Journal of Geriatric Psychiatry* **26**, 239–246.

Goldberg DP (1972). *The Detection of Psychiatric Illness by Questionnaire*. Oxford University Press: Oxford.

Goldberg DP, Hillier VF (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine* **9**, 139–145.

Goldberg DP, Williams P (1988). *A User's Guide to the General Health Questionnaire*. NFER-Nelson: Windsor.

Haro JM, Arbabzadeh-Bouchez S, Brugha TS, de Girolamo G, Guyer ME, Jin R, Lepine JP, Mazzi F, Reneses B, Vilagut G, Sampson NA, Kessler RC (2006). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *International Journal of Methods in Psychiatric Research* **15**, 167–180.

Jordanova V, Wickramesinghe C, Gerada C, Prince M (2004). Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. *Psychological Medicine* **34**, 1013–1024.

Klinkman MS, Coyne JC, Gallo S, Schwenk TL (1997). Can case-finding instruments be used to improve physician detection of depression in primary care? *Archives of Family Medicine* **6**, 567–573.

Lewis G (1994). Assessing psychiatric disorder with a human interviewer or a computer. *Journal of Epidemiology and Community Health* **48**, 207–210.

Lewis G, Pelosi AJ, Araya R, Dunn G (1992). Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine* **22**, 465–486.

Lewis G, Pelosi AJ, Glover E, Wilkinson G, Stansfeld SA, Williams P, Shepherd M (1988). The development of a computerised assessment for minor psychiatric disorder. *Psychological Medicine* **18**, 737–745.

Marmot MG, Davey Smith G, Stansfeld SA, Patel C, North F, Head J, White I, Brunner EJ, Feeney A (1991). Health inequalities among British civil servants: the Whitehall II study. *Lancet* **337**, 1387–1393.

Metz CE (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298.

Patton GC, Coffey C, Posterino M, Carlin JB, Wolfe R, Bowes G (1999). A computerised screening instrument for adolescent depression: population-based validation and application to a two-phase case-control study. *Social Psychiatry and Psychiatric Epidemiology* **34**, 166–172.

Peters L, Clark D, Carroll F (1998). Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders. *Psychological Medicine* **28**, 893–901.

Pez O, Gilbert F, Bitfoi A, Carta MG, Jordanova V, Garcia-Mahia C, Mateos-Alvarez R, Prince M, Tudorache B, Blatier C, Kovess-Masfety V (2010). Validity across translations of short survey psychiatric diagnostic instruments: CIDI-SF and CIS-R versus SCID-I/NP in four European countries. *Social Psychiatry and Psychiatric Epidemiology* **45**, 1149–1159.

Radloff LS (1977). The CES-D Scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385–401.

Stahl D, Sum CF, Lum SS, Liow PH, Chan YH, Verma S, Chua HC, Chong SA (2008). Screening for depressive symptoms: validation of the Center for Epidemiologic Studies Depression Scale (CES-D) in a multiethnic group of patients with diabetes in Singapore. *Diabetes Care* **31**, 1118–1119.

Stansfeld S, Head J, Bartley M, Fonagy P (2008). Social position, early deprivation and the development of attachment. *Social Psychiatry and Psychiatric Epidemiology* **43**, 516–526.

Stansfeld SA, Head J, Marmot MG (1998). Explaining social class differences in depression and well-being. *Social Psychiatry and Psychiatric Epidemiology* **33**, 1–9.

Stansfeld SA, Marmot MG (1992). Social class and minor psychiatric disorder in British civil servants: a validated screening survey using the General Health Questionnaire. *Psychological Medicine* **22**, 739–749.

Thomas JL, Jones GN, Scarinci IC, Mehan DJ, Brantley PJ (2001). The utility of the CES-D as a depression screening measure among low-income women attending primary care clinics. The Center for Epidemiologic Studies-Depression. *International Journal of Psychiatry in Medicine* **31**, 25–40.

Williams Jr. JW, Noel PH, Cordes JA, Ramirez G, Pignone M (2002). Is this patient clinically depressed? *Journal of the American Medical Association* **287**, 1160–1170.

Wittchen HU (1994). Reliability and validity studies of the WHO – Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.